An Iterative Coordinate Descent Algorithm for High-Dimensional Nonconvex Penalized Quantile Regression

Bo PENG and Lan WANG

We propose and study a new iterative coordinate descent algorithm (QICD) for solving nonconvex penalized quantile regression in high dimension. By permitting different subsets of covariates to be relevant for modeling the response variable at different quantiles, nonconvex penalized quantile regression provides a flexible approach for modeling high-dimensional data with heterogeneity. Although its theory has been investigated recently, its computation remains highly challenging when p is large due to the nonsmoothness of the quantile loss function and the nonconvexity of the penalty function. Existing coordinate descent algorithms for penalized least-squares regression cannot be directly applied. We establish the convergence property of the proposed algorithm under some regularity conditions for a general class of nonconvex penalty functions including popular choices such as SCAD (smoothly clipped absolute deviation) and MCP (minimax concave penalty). Our Monte Carlo study confirms that QICD substantially improves the computational speed in the $p \gg n$ setting. We illustrate the application by analyzing a microarray dataset.

Key Words: Coordinate descent quantile regression; High-dimensional data; MCP; Nonconvex penalty; SCAD; Variable selection.

1. INTRODUCTION

We consider a high-dimensional regression setting where the number of predictors p greatly exceeds the sample size n. The random sample $\{Y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$, is assumed to arise from the model $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$, where $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ is a $(p+1)$-dimensional vector of covariates with $x_{i0} = 1$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denotes the vector of parameters, and ϵ_i is the random error. The parameter $\boldsymbol{\beta}$ is assumed to be sparse in the sense that most of its components are zero. We are interested in identifying and estimating the nonzero components of $\boldsymbol{\beta}$ when $p \gg n$. For this problem, a large amount of literature in recent years have been devoted to the penalized least-squares regression approach, which

Bo Peng (E-mail: *peng0199@umn.edu*) is a graduate student and Lan Wang (E-mail: *wangx346@umn.edu*) is Associate Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455.

Journal of Computational and Graphical Statistics, Volume 00, Number 0, Pages 1–20
DOI: 10.1080/10618600.2014.913516

minimizes

$$n^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{i=1}^{p} p_\lambda(|\beta_j|),$$

with $p_\lambda(\cdot)$ denoting a penalty function with a tuning parameter λ. We refer to the recent survey papers by Bickel and Li (2006) and Fan and Lv (2010) for many relevant references.

Although the penalized linear squares approach is useful, it only focuses on the central tendency of the conditional distribution. A certain covariate may not have significant influence on the mean value of the response variable but may have a strong effect at the upper quantile of the conditional distribution due to the heterogeneous nature of the data. It is also likely that a certain covariate may have different effects at different segments of the conditional distribution.

A useful alternative approach is to use penalized quantile regression, which enjoys several appealing features for analyzing data with heterogeneity due to either heteroscedastic variance or other forms of non-location-scale covariate effects. The penalized quantile regression framework assumes that only a small number of covariates influence the conditional distribution of the response variable given the high-dimensional covariates; however, the sets of relevant covariates may vary when we consider different segments of the conditional distribution. This general interpretation of sparsity is more realistic for real data problems. By considering different quantiles, we can explore the entire conditional distribution of the response variable and obtain a more realistic picture of the sparsity pattern. Furthermore, in many real applications, the conditional quantiles, such as the conditional median, are of direct interest to the researchers. And penalized quantile regression is naturally robust against outliers in the response space.

In the setting $p \gg n$, the theory for penalized quantile regression has been only recently systematically studied by Belloni and Chernozhukov (2011) for the Lasso penalty and by Wang, Wu, and Li (2012) for the nonconvex penalties such as SCAD (smoothly clipped absolute deviation, Fan and Li 2001) or MCP (minimax concave penalty, Zhang 2010). It is also noteworthy that in the fixed p case, the theory for penalized quantile regression was investigated by Zou and Yuan (2008), Wu and Liu (2009), Kai, Li, and Zou (2011), among others.

For unpenalized quantile regression, Koenker and Park (1996) proposed a useful interior point algorithm, and Hunter and Lange (2000) developed an effective MM algorithm which majorizes the quantile loss function by a quadratic function. Several algorithms have also been developed for penalized quantile regression. For Lasso penalized quantile regression, Li and Zhu (2008) proposed an algorithm that computes the entire solution path; Wu and Lange's work (2008) includes a fast greedy coordinate descent algorithm for median regression. However, neither algorithm applies to nonconvex penalties. A linear programming based modified local linear algorithm (LLA) (Zou and Li 2008) was used in Wang, Wu, and Li (2012) for nonconvex penalized quantile regression, but its computation is noticeably slow when p is large. Moreover, the aforementioned work have not studied the algorithm convergence theory.

In this article, we focus on nonconvex penalized quantile regression. The nonconvex penalty is known to alleviate the bias of Lasso and lead to consistent variable selection under weaker conditions (e.g., Fan and Li 2001). To tackle the computational challenges

caused by the nonsmooth quantile loss function and the nonconvex penalty function when p is large, we propose a new iterative coordinate descent algorithm and study its convergence property. The new algorithm achieves fast computation by successively solving a sequence of univariate minimization subproblems. It combines the idea of the MM algorithm (e.g., Hunter and Lange 2004; Lange 2004; Hunter and Li 2005) with that of the coordinate descent algorithm. We refer to this new iterative coordinate descent algorithm as QICD, where Q stands for quantile. We consider a general class of nonconvex penalty functions and establish the convergence property of the QICD algorithm by extending Tseng's (2001) theory for the convergence of the coordinate descent algorithm. It is noted that Tseng (2001) required a quasiconvexity condition, which is not met by nonconvex penalized quantile regression.

The coordinate descent algorithm was systematically investigated for convex problems, such as Lasso, in the independent work of Friedman et al. (2007) and Wu and Lange (2008), the idea of which can be traced back to Fu (1998) and Daubechies, Defrise, and De Mol (2004). For nonconvex penalized least squares regression, coordinate descent algorithms and their convergence theory were recently investigated by Breheny and Huang (2011) and Mazumder, Friedman, and Hastie (2011). Jiang and Huang (2012) proposed a new coordinate descent algorithm for nonconvex penalized generalized linear models which enjoys the appealing property of avoiding the computation of a scaling factor in each update of the solutions. These algorithms are very effective in large-scale problems but do not apply to nonconvex penalized quantile regression.

In Section 2, we introduce nonconvex penalized quantile regression. Section 3 describes the QICD algorithm. In Section 4, we establish the convergence property of the QICD algorithm. In Section 5, we investigate the performance of the QICD algorithm through Monte Carlo studies and demonstrate its application to a real data example. The technical details are presented in the Appendix.

2. NONCONVEX PENALIZED QUANTILE REGRESSION

Quantile regression was proposed in the seminal work of Koenker and Bassett (1978). Let the conditional distribution function of Y_i given $\mathbf{x}_i$ be $F(y|\mathbf{x}) = P(Y_i \le y|\mathbf{x}_i = \mathbf{x})$. For $0 < \tau < 1$, the τth conditional quantile of Y_i given $\mathbf{x}_i = \mathbf{x}$ is defined as $Q_{Y|\mathbf{x}}(\tau) = \inf\{t : F(t|\mathbf{x}) \ge \tau\}$. The case $\tau = 1/2$ corresponds to the conditional median. A linear quantile regression model assumes $Q_{Y|\mathbf{x}}(\tau) = \mathbf{x}^T\boldsymbol{\beta}$ for an unknown parameter vector of parameters $\boldsymbol{\beta}$. We refer to Koenker (2005) for a comprehensive introduction to unpenalized quantile regression.

In the high-dimensional setting, most of the components in $\boldsymbol{\beta}$ are zero under the sparsity assumption. The penalized quantile regression estimator for $\boldsymbol{\beta}$ is obtained by minimizing

$$Q(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function. The tuning parameter λ in the penalty function $p_\lambda(\cdot)$ controls the model complexity. Furthermore, $p_\lambda(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $p'_\lambda(t)$ on

$(0, +\infty)$. In this article, we consider a general class of nonconvex penalty functions, which in particular includes the two popular choices: SCAD and MCP. The SCAD penalty function (Fan and Li 2001) is defined by

$$p_\lambda(|\beta|) = \lambda|\beta| I(0 \le |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \le |\beta| \le a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda),$$

for some $a > 2$; while the MCP penalty function (Zhang 2010) has the form

$$p_\lambda(|\beta|) = \lambda\Big(|\beta| - \frac{\beta^2}{2a\lambda}\Big) I(0 \le |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \ge a\lambda),$$

for some $a > 1$. Both penalty functions are singular at the origin to achieve sparsity of estimation. They also both remain constant when $|\beta|$ exceeds $a\lambda$, which avoids over-penalizing large coefficients and alleviates the bias problem associated with Lasso. See Figure 1 for an illustration.

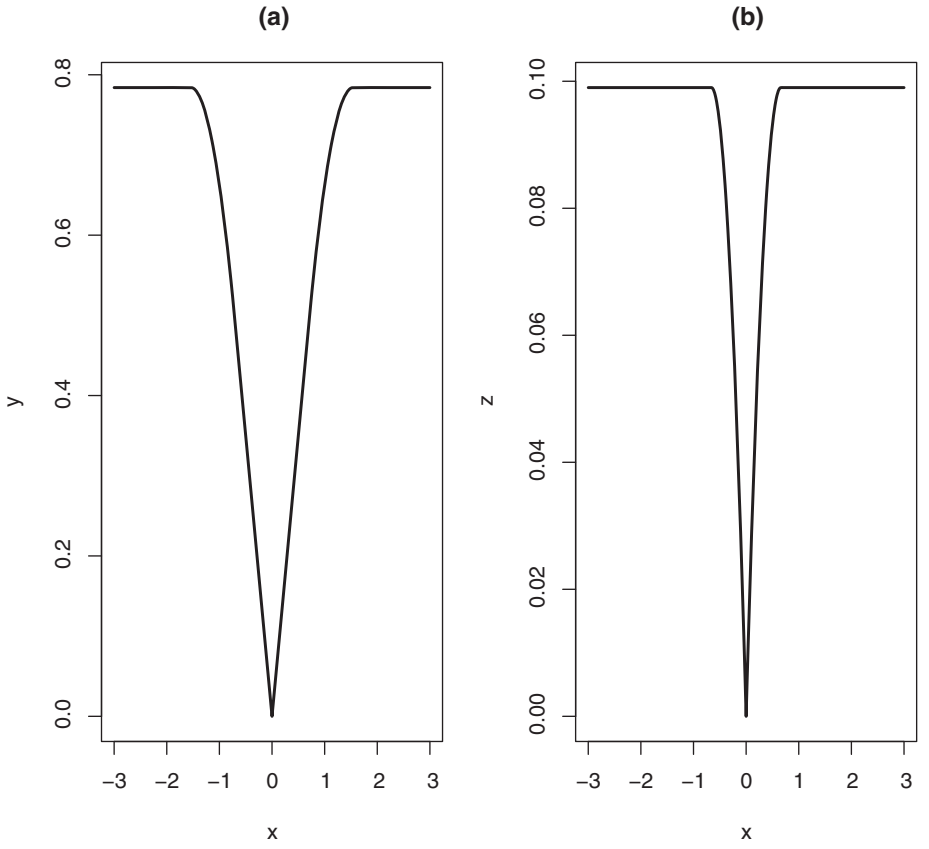

Figure 1. (a) SCAD penalty function with $\lambda = 0.7$, $a = 2.2$; (b) MCP penalty function with $\lambda = 0.3$, $a = 2.2$.

Wang, Wu, and Li (2012) recently established the asymptotic theory for nonconvex penalized sparse quantile regression in the ultrahigh dimension, where p is allowed to grow at an exponential rate of n. The rate depends on the dimension of the true model and the magnitude of the smallest signal (see the remark following Theorem 2.4 in Wang, Wu, and Li 2012). Their work made use of a novel sufficient optimality condition which relies on a convex differencing representation of the penalized quantile loss function. Their theory requires much weaker assumptions than those in the literature for penalized least squares regression. In particular, there is no need to impose restrictive distributional or moment conditions on the random errors.

To solve the minimization problem in (1), Wang, Wu, and Li (2012) proposed a linear programming based modification of the LLA algorithm (Zou and Li 2008). More specifically, starting with initial values $\tilde{\beta}_j^{(0)} = 0$ for $j = 1, 2, \ldots, p$. At the tth step ($t > 1$), given $\tilde{\boldsymbol{\beta}}^{(t-1)} = (\tilde{\beta}_1^{(t-1)}, \ldots, \tilde{\beta}_p^{(t-1)})^T$, the update $\tilde{\boldsymbol{\beta}}^{(t)}$ is obtained by minimizing

$$n^{-1} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right) + \sum_{j=1}^{p} w_j^{(t-1)} |\beta_j|, \tag{2}$$

where $w_j^{(t-1)} = p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)$. Following the literature, when $\tilde{\beta}_j^{(t-1)} = 0$, we take $p'_\lambda(0)$ as $p'_\lambda(0+) = \lambda$. With the aid of slack variables ξ_i^+, ξ_i^-, and ζ_j, the convex optimization problem in (2) can be equivalently written as

$$\min_{\boldsymbol{\xi}, \boldsymbol{\zeta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tau \xi_i^+ + (1 - \tau) \xi_i^-) + \sum_{j=1}^{p} w_j^{(t-1)} \zeta_j \right\} \tag{3}$$

subject to

$$\begin{aligned} &\xi_i^+ - \xi_i^- = Y_i - \mathbf{x}_i^T \boldsymbol{\beta};\ i = 1, 2, \ldots, n, \\ &\xi_i^+ \geq 0, \xi_i^- \geq 0;\ i = 1, 2, \ldots, n, \\ &\zeta_j \geq \beta_j, \zeta_j \geq -\beta_j;\ j = 1, 2, \ldots, p. \end{aligned}$$

Note that (3) is a linear programming problem and can be solved using many existing software packages. The convergence theory of this algorithm has not been investigated. Our numerical experience (Section 5) indicates that the proposed QICD algorithm is substantially faster than the above algorithm, which we refer to as the LLA algorithm in the sequel.

3. THE QICD ALGORITHM

The QICD algorithm combines the idea of the MM algorithm with that of the coordinate descent algorithm. More specifically, we first replace the nonconvex penalty function by its majorization function to create a surrogate objective function. Then, we minimize the surrogate objective function with respect to a single parameter each time and cycle through all parameters until convergence. For each coordinate descent step, we only need to compute a univariate weighted median, which ensures fast computation.

3.1 THE MAJORIZATION MINIMIZATION STEP

We consider a *majorization function* $\phi_{\beta_0}(\beta)$, which majorizes $p_\lambda(|\beta|)$ at β_0 in the sense that

$$\phi_{\beta_0}(\beta) \geq p_\lambda(|\beta|) \quad \text{for all } \beta, \text{ with equality when } \beta = \beta_0. \tag{4}$$

Let $\boldsymbol{\beta}^{(k)}$ denote the value of $\boldsymbol{\beta}$ after the kth iteration, $k = 1, 2, \ldots$. Let $p'_\lambda(|\beta|+)$ denotes the limit of $p'_\lambda(x)$ as $x \to |\beta|$ from the above. Furthermore, we assume that $p_\lambda(\cdot)$ is piecewise differentiable so that $p'_\lambda(|\beta|+)$ exists for all β. Then, in the kth iteration,

$$\phi_{\beta_j^{(k-1)}}(|\beta_j|) = p'_\lambda\left(|\beta_j^{(k-1)}|+\right)|\beta_j| - p'_\lambda\left(|\beta_j^{(k-1)}|+\right)|\beta_j^{(k-1)}| + p_\lambda\left(|\beta_j^{(k-1)}|\right) \tag{5}$$

majorizes the penalty function $p_\lambda(|\beta_j|)$, $k = 1, 2, \ldots; j = 1, 2, \ldots, p$; that is,

$$\phi_{\beta_j^{(k-1)}}(|\beta_j|) \geq p_\lambda(|\beta_j|) \text{ for all } \beta_j, \text{ with equality when } \beta_j = \beta_j^{(k-1)}, \tag{6}$$

see Lemma 1. Figure 2 depicts the SCAD penalty function and its majorization function; the figure for the MCP penalty looks similar and is omitted.

Subsequently, the penalized objective function $Q(\boldsymbol{\beta})$ defined in (1) is majorized by

$$Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) + \sum_{j=1}^{p}\phi_{\beta_j^{(k-1)}}(|\beta_j|) \tag{7}$$

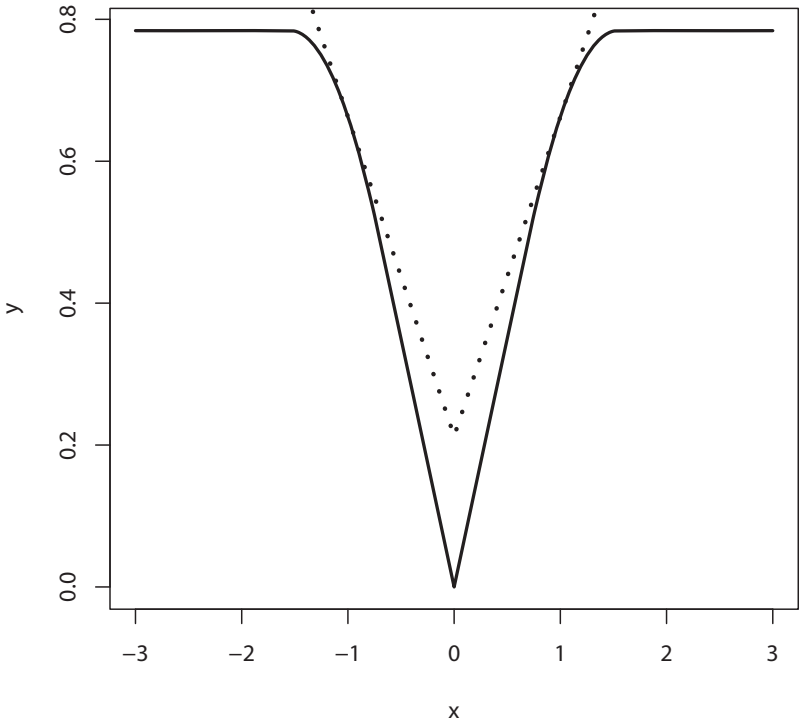

Figure 2. SCAD penalty function (solid line) and its majorization function (dotted line) $\phi_{\beta_0}(\beta)$ with $\lambda = 0.7$, $a = 2.2$.

at the kth iteration. It can be shown that any decrease of the value of $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ results in a decrease of the value of $Q(\boldsymbol{\beta})$. Hence, we minimize the majorization function $Q_{\boldsymbol{\beta}^{k-1}}(\boldsymbol{\beta})$ at iteration k to update the value of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(k)} = \arg\min_{\boldsymbol{\beta}} Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}) \tag{8}$$

The above iterative scheme decreases the value of $Q(\boldsymbol{\beta})$ monotonically in each iteration. This property is summarized in Lemma 1. We note that when considering nonconvex penalized generalized linear models, Jiang and Huang (2012) applied a majorization step on the loss function to avoid the computation of a scaling factor in each update of the solution. Different from their approach, our majorization step is applied to the nonconvex penalty function. The majorization step solves two problems at once. First, it transforms the problem of minimizing a nonconvex objective function to a sequence of convex minimization problems, for which the coordinate descent algorithm can be applied. Second, the majorized penalized quantile loss function is quasiconvex, which allows us to apply the results in Tseng (2001) to further study the convergence property of the proposed algorithm.

3.2 THE COORDINATE DESCENT STEP

To solve the minimization problem in (8), we employ the idea of the "one-at-a-time" coordinate descent algorithm; that is, to update the jth coordinate, we treat the other coordinates as fixed. The subiteration of the coordinate descent algorithm is incorporated within each iteration of the majorization minimization step.

Assume that at the beginning of the kth iteration, the value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{(k-1)}$. To minimize $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$, we apply the coordinate descent algorithm, which cycles through all the $(p+1)$ covariates at each subiteration . Consider the rth subiteration of the kth iteration. Suppose we have finished updating the estimates of the coefficients for $x_{i0}, x_{i1}, \ldots, x_{i(j-1)}$ and obtained $\boldsymbol{\beta}_{j-1}^{(k)(r)} = (\beta_0^{(k)(r+1)}, \ldots, \beta_{j-1}^{(k)(r+1)}, \beta_j^{(k)(r)}, \ldots, \beta_p^{(k)(r)})^T$. Next, we update the estimate for the coefficient of x_{ij} by

$$\begin{aligned}
\beta_j^{(k)(r+1)} &= \arg\min_{\beta_j} Q_{\boldsymbol{\beta}^{(k-1)}}\left(\boldsymbol{\beta}_{j-1}^{(k)(r)}\right) \\
&= \arg\min_{\beta_j} \left\{ n^{-1} \left[\sum_{i=1}^{n} \rho_\tau \left(Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - x_{ij}\beta_j \right.\right.\right. \\
&\quad \left.\left.\left. - \sum_{s>j} x_{is}\beta_s^{(k)(r)} \right) \right] + \left[\sum_{s<j} \phi_{\beta_s^{(k-1)}}\left(|\beta_s^{(k)(r+1)}|\right) + \phi_{\beta_j^{(k-1)}}(|\beta_j|) \right.\right. \\
&\quad \left.\left. + \sum_{s>j} \phi_{\beta_s^{(k-1)}}\left(|\beta_s^{(k)(r)}|\right) \right] \right\} \\
&= \arg\min_{\beta_j} \left\{ n^{-1} \left[\sum_{i=1}^{n} \rho_\tau \left(Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - x_{ij}\beta_j \right.\right.\right.
\end{aligned}$$

$$- \sum_{s>j} x_{is}\beta_s^{(k)(r)}\Bigg)\Bigg] + p'_\lambda\left(|\beta_j^{(k-1)}|+\right)|\beta_j|\Bigg\}. \tag{9}$$

In the above minimization, $\boldsymbol{\beta}^{(k-1)}$ and all the other coordinates in $\boldsymbol{\beta}_{j-1}^{(k)(r)}$ are held fixed. An important observation is that (9) can be equivalently expressed as a minimization problem for weighted median regression. To see the connection, we rewrite (9) as

$$\min_{\beta_j}\left\{(n+1)^{-1}\sum_{i=1}^{n+1}\omega_{ij}|u_{ij}|\right\}, \tag{10}$$

where

$$u_{ij} = \begin{cases} \dfrac{Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - \sum_{s>j} x_{is}\beta_s^{(k)(r)}}{x_{ij}} - \beta_j, & i = 1, 2, \ldots, n, \\ \beta_j, & i = n+1, \end{cases}$$

and

$$\omega_{ij} = \begin{cases} n^{-1}|x_{ij}(\tau - I(u_{ij}x_{ij} < 0))|, & i = 1, 2, \ldots, n, \\ p'_\lambda(|\beta_j^{(k-1)}|+), & i = n+1. \end{cases}$$

Therefore, $\beta_j^{(k)(r+1)}$can be obtained by solving a weighted median problem using the above $n+1$ pseudo-observations, $j > 1$. When $j = 0$, $\beta_0^{(k)(r+1)}$ can be calculated by using only n pseudo-observations since no penalty is given to $\beta_0^{(k)(r+1)}$. A similar observation was made for the Lasso penalized median regression by Wu and Lange (2008). In our algorithm, we calculate the weighted median by *quicksort*, also known as partition-exchange sort, which ensures the high speed for updating $\beta_j^{(k)(r+1)}$.

The above computation yields

$$\boldsymbol{\beta}_j^{(k)(r)} = (\beta_0^{(k)(r+1)}, \ldots, \beta_j^{(k)(r+1)}, \beta_{j+1}^{(k)(r)}, \ldots, \beta_p^{(k)(r)})^T.$$

This process is repeated for $r = 1, 2, \ldots,$ until convergence. Then, we update $\boldsymbol{\beta}^{(k-1)}$ to $\boldsymbol{\beta}^{(k)}$.

3.3 CHOICE OF THE TUNING PARAMETER

Algorithm 3.1 summarizes the details of the QICD algorithm for a given tuning parameter λ. In real applications, the choice of λ is important. Cross-validation is popular but is

observed to often result in overfitting (Wang, Li, and Tsai 2007). Moreover, cross-validation is time-consuming when p is notably large.

Algorithm 3.1: QICD ALGORITHM(k, r, j, p, $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(k)}$)

comment: Input an initial value $\boldsymbol{\beta}^{(0)}$

for $k \geq 0$

then
- **repeat**
- **for** $r \geq 0$
- **then**
  - **for** $j \in \{0, 1, 2, \ldots, p\}$
  - **repeat**
    - **comment:** Calculate the weighted median in (10).
    - $\boldsymbol{\beta}_{j+1}^{(k)(r)} \leftarrow \boldsymbol{\beta}_{j}^{(k)(r)}$
    - **for** $j = p$
      - **then** $r \leftarrow r + 1$
    - $j \leftarrow j + 1 \pmod{p}$
  - **until** $\boldsymbol{\beta}_{j}^{(k)(r)}$ converge to $\boldsymbol{\beta}^{*}$
  - $\boldsymbol{\beta}^{(k)} \leftarrow \boldsymbol{\beta}^{*}$
- **until** $\boldsymbol{\beta}^{(k)}$ converge to $\widehat{\boldsymbol{\beta}}$

then return $(\widehat{\boldsymbol{\beta}})$

High-dimensional BIC-type criterion for nonconvex penalized least-squares regression with diverging p was recently investigated by Wang, Li, and Leng (2009), Chen and Chen (2008), Kim, Kwon, and Choi (2012), and Wang, Kim, and Li (2013), among others. Lee, Noh, and Park (2013) recently proposed high-dimensional BIC for quantile regression when p is much larger than n. Motivated by their work, we consider the following high-dimensional BIC criterion. Let $\boldsymbol{\beta}_{\lambda} = (\beta_{\lambda,1}, ..., \beta_{\lambda,p})^T$ be the penalized estimator obtained with the tuning parameter λ, and let $\mathcal{S}_{\lambda} \equiv \{j : \beta_{\lambda,j} \neq 0, 1 \leq j \leq p\}$ be the index set of nonzero coefficients selected by this estimator. Define

$$\mathrm{HBIC}(\lambda) = \log\left(\sum_{i=1}^{n} \rho_{\tau}(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{\lambda})\right) + |\mathcal{S}_{\lambda}| \frac{\log(\log n)}{n} C_n, \tag{11}$$

where $|\mathcal{S}_{\lambda}|$ is the cardinality of the set $\mathcal{S}_{\lambda}$, and C_n is a sequence of positive constants diverging to infinity as n increases. We select the value of λ that minimizes HBIC(λ). In practice, we recommend to take $C_n = O(\log(p))$, which we find to work well in a variety of settings.

4. CONVERGENCE THEORY

The main result in this section establishes that under some regularity conditions, the proposed QICD algorithm converges to a stationary point of the penalized objective function in (1).

Lemma 1 summarizes an important property of the majorization minimization step.

Lemma 1. Assume that the penalty function $p_\lambda(\cdot)$ in (1) is piecewise differentiable, nondecreasing and concave on $(0, \infty)$, and $p_\lambda(\cdot)$ is continuous at 0 with $p'_\lambda(0+) < \infty$. Then,

1. the function $\phi_{\beta_j^{(k-1)}}(\beta)$ defined in (5) majorizes $p_\lambda(|\beta|)$ at the points $\pm|\beta_j^{(k-1)}|$;
2. the function $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ defined in (7) majorizes $Q(\boldsymbol{\beta})$ at the points $\pm|\boldsymbol{\beta}^{(k-1)}|$;
3. the majorization minimization step has the descent property, that is, for all $k = 1, 2, \ldots$

$$Q(\boldsymbol{\beta}^{(k)}) \leq Q(\boldsymbol{\beta}^{(k-1)}). \tag{12}$$

The general theory of Tseng (2001) on the coordinate descent algorithm does not apply to the penalized quantile objective function in (1). This is because nonconvex penalized quantile regression does not meet the *quasiconvex* condition. Lemma 2 suggests that if we consider the majorized loss function $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$, then the coordinate descent step yields a coordinate-wise minimum (see Appendix A for the definition).

Lemma 2. Assume $p_\lambda(\cdot)$ satisfies the conditions in Lemma 1. Then, the $\boldsymbol{\beta}^{(k)}$ defined in Section 3.2 is a coordinate-wise minimum point of $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$, $k = 1, 2, \ldots$.

Lemma 3 describes the convergence behavior of $Q(\boldsymbol{\beta}^{(k)})$. It indicates that $Q(\boldsymbol{\beta}^{(k)})$ follows similar convergence behavior as its majorization function $Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)})$ under some weak conditions.

Lemma 3. If $Q(\boldsymbol{\beta}^{(0)}) < +\infty$, then $\{Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)})\}$ is a bounded and decreasing sequence with respect to k. If we denote $\lim_{k\to\infty} Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)})$ by A, then $Q(\boldsymbol{\beta}^*) = A$, where $\boldsymbol{\beta}^*$ be an arbitrary cluster point of $\{\boldsymbol{\beta}^{(k)}\}$.

The convergence property of the QICD algorithm is established by combining the results of the two preceding lemmas and using a result in Bazaraa, Sherali, and Shetty (2006, see Lemma B.1 in Appendix B). Theorem 1 states that every cluster point of the QICD algorithm is a stationary point of the penalized quantile loss function $Q(\boldsymbol{\beta})$.

Theorem 1. (Convergence property of QICD) Consider the penalized quantile loss function $Q(\boldsymbol{\beta})$ in (1), where the given data $(Y, \mathbf{X})$ lie on a compact set and $Q(\boldsymbol{\beta}^{(0)}) < +\infty$ for an initial value $\boldsymbol{\beta}^{(0)}$. Suppose that the penalty function $p_\lambda(\cdot)$ satisfies the conditions in Lemma 1 and $p'_\lambda(|\theta|+) = p'_\lambda(|\theta|-)$ on $(0, \infty)$. Consider an arbitrary cluster point $\boldsymbol{\beta}^{**}$ of $\{\boldsymbol{\beta}^{(k-1)}\}$, that is, there exists a sequence $\{k_m\}$ such that $\lim_{m\to\infty} \boldsymbol{\beta}^{(k_m-1)} = \boldsymbol{\beta}^{**}$. Let $\boldsymbol{\beta}^*$ be an arbitrary cluster point of $\{\boldsymbol{\beta}^{(k_m)}\}$. Assume $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ is *regular* at $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$. Then, $\boldsymbol{\beta}^{**}$ is a stationary point of $Q(\boldsymbol{\beta})$. In particular, every cluster point of the sequence generated by the QICD algorithm $\{\boldsymbol{\beta}^{(k)}\}$ is a *stationary point* of $Q(\boldsymbol{\beta})$.

Note that the condition on $(Y, \mathbf{X})$ is a mild assumption. And the conditions on the penalty functions are satisfied by the popular nonconvex SCAD and MCP penalties. The proof of Theorem 1 is provided in Appendix B.

5. NUMERICAL EXAMPLES

5.1 MONTE CARLO SIMULATIONS

We first generate $(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_p)^T$ from the multivariate normal distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\sigma_{jk})_{p\times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. Then, we set $X_1 = \Phi(\tilde{X}_1)$ and $X_j = \tilde{X}_j$ for $j = 2, 3, \ldots, p$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Then, we generate the response variable from the following regression model:

$$Y = X_6 + X_{12} + X_{15} + X_{20} + 0.7X_1\epsilon$$

where the random error $\epsilon \sim N(0, 1)$ is independent of the covariates. In this model, the τth conditional quantile function is $X_6 + X_{12} + X_{15} + X_{20} + 0.7X_1\Phi^{-1}(\tau)$, where $\Phi^{-1}(\tau)$ denotes the τth quantile of the standard normal distribution. Hence, X_1 does not influence the center of the conditional distribution, but plays an important role when considering other conditional quantiles.

We consider sample size $n = 300$, $p = 1000$ and 2000, and three different quantiles $\tau = 0.3, 0.5$ and 0.7. For each simulation scenario, we perform 100 simulation runs. As the comparison of Lasso penalized quantile regression and nonconvex penalized quantile regression has been considered in Wang, Wu, and Li (2012), we focus on comparing the performance of the QICD with that of LLA, with SCAD and MCP penalty functions. For both procedures, the HBIC discussed in Section 3.3 is applied to choose the tuning parameter.

The convergence criteria used in implementing the QICD algorithm are as follows: (i) the coordinate descent step in each iteration stops if the absolute difference of the successive subiterations is less than 10^{-6} (convergence of coefficients in subiteration) and the number of subiterations exceeds p; (ii) the majoriaztion minimization step stops if the absolute difference of the successive iterations is less than 10^{-6} (convergence of coefficients in iteration) or the number of iterations exceeds 100.

We evaluate the two algorithms by the estimation error, the model selection ability and the computational speed. More specifically, for a given estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$, we consider the following five criteria:

Size : the average number of nonzero regression coefficients $\widehat{\beta}_j \neq 0$ for $j = 1, 2, \ldots, p$;

P1: the proportion of simulation runs X_6, X_{12}, $X_{15,}$ and X_{20} are selected.

P2 : the proportion of simulation runs X_1 is selected.

AE : the absolute estimation error defined by $\sum_{j=0}^{p} |\widehat{\beta}_j - \beta_j|$.

Time : the running times (CPU seconds) for each method in each repetition (the process of calculating the estimate $\widehat{\boldsymbol{\beta}}$).

Tables 1–2 summarize the simulation results for $p = 1000$ and 2000, respectively. We observe that both the QICD algorithm and the LLA algorithm have satisfactory performance in terms of estimation error and model selection accuracy. The QICD algorithm

Table 1. Simulation results ($p = 1000$)

Method	Size	P1	P2	AE	Time
QICD-SCAD ($\tau = 0.5$)	5.15 (0.51)	100%	0%	0.04 (0.01)	1.53 (0.40)
QICD-SCAD ($\tau = 0.3$)	7.53 (2.11)	100%	94%	0.11 (0.02)	1.57 (0.39)
QICD-SCAD ($\tau = 0.7$)	8.02 (2.42)	100%	93%	0.11 (0.03)	1.56 (0.33)
LLA-SCAD ($\tau = 0.5$)	5.00 (0.00)	100%	0%	0.04 (0.01)	24.78 (1.54)
LLA-SCAD ($\tau = 0.3$)	7.68 (1.44)	100%	91%	0.11 (0.03)	39.43 (3.27)
LLA-SCAD ($\tau = 0.7$)	10.86 (2.06)	100%	94%	0.13 (0.02)	28.59 (1.94)
QICD-MCP ($\tau = 0.5$)	5.25 (0.72)	100%	0%	0.04 (0.01)	1.52 (0.41)
QICD-MCP ($\tau = 0.3$)	7.57 (1.95)	100%	96%	0.12 (0.03)	1.94 (0.60)
QICD-MCP ($\tau = 0.7$)	8.40 (2.78)	100%	96%	0.12 (0.03)	1.77 (0.36)
LLA-MCP ($\tau = 0.5$)	5.00 (0.00)	100%	0%	0.04 (0.01)	29.88 (6.92)
LLA-MCP ($\tau = 0.3$)	8.58 (1.68)	100%	93%	0.12 (0.03)	38.54 (8.50)
LLA-MCP ($\tau = 0.7$)	9.89 (1.81)	100%	95 %	0.12 (0.02)	69.69 (19.70)

is remarkably faster than the LLA algorithm. For $p = 1000$, the QICD algorithm takes less than 2 sec on average to finish a repetition, which is $\frac{1}{15}$ of the runtime of the average LLA algorithm; for $p = 2000$ case, the LLA usually needs 100 sec or more to finish one repetition, which is much longer than the runtime of the QICD algorithm (under 4 sec on average). Furthermore, the LLA algorithm displays great variation on running time; especially in the case $p = 2000$ where the standard deviation of then running time when $\tau = 0.7$ could be as large as 80 sec. In contrast, the QICD algorithm is much more stable with the standard deviation of the running time less than 1.5 sec.

It is also observed that the QICD algorithm tends to select a sparser model but with comparable estimation error comparing with the LLA algorithm. In particular, for $p = 2000$, the estimation error associated the QICD algorithm is slightly smaller than that of the LLA algorithm; meanwhile, the size of the nonezero coefficients (around 5) for the QICD algorithm when $\tau = 0.5$ is also notably smaller than that of the LLA algorithm (around 13). In summary, the fast computation of the QICD algorithm does not come at the cost of sacrificing its performance.

Table 2. Simulation results ($p = 2000$)

Method	Size	P1	P2	AE	Time
QICD-SCAD ($\tau = 0.5$)	5.23 (0.88)	100%	0%	0.04 (0.01)	3.37 (1.08)
QICD-SCAD ($\tau = 0.3$)	8.00 (2.49)	100%	93%	0.11 (0.03)	2.96 (0.76)
QICD-SCAD ($\tau = 0.7$)	8.52 (2.16)	100%	93%	0.12 (0.03)	3.19 (0.72)
LLA-SCAD ($\tau = 0.5$)	13.48 (2.93)	100%	0%	0.07 (0.02)	235.17 (17.57)
LLA-SCAD ($\tau = 0.3$)	8.41 (1.68)	100%	87%	0.11 (0.03)	148.94 (8.00)
LLA-SCAD ($\tau = 0.7$)	9.67 (2.21)	100%	92%	0.12 (0.03)	214.23 (20.21)
QICD-SCAD ($\tau = 0.5$)	5.33 (1.18)	100%	0%	0.04 (0.02)	3.05 (0.80)
QICD-SCAD ($\tau = 0.3$)	8.21 (2.72)	100%	92%	0.12 (0.03)	3.20 (0.76)
QICD-SCAD ($\tau = 0.7$)	8.48 (2.17)	100%	93%	0.12 (0.03)	3.72 (1.05)
LLA-MCP ($\tau = 0.5$)	13.67 (2.97)	100%	0%	0.07 (0.02)	166.12 (38.98)
LLA-MCP ($\tau = 0.3$)	8.45 (2.03)	100%	88%	0.12 (0.03)	219.03 (46.37)
LLA-MCP ($\tau = 0.7$)	8.65 (1.77)	100%	92 %	0.12 (0.02)	354.01 (83.76)

5.2 AN APPLICATION

We next analyze the microarray dataset of Scheetz et al. (2006) for studying expression quantitative trait locus (eQTL) mapping in the laboratory rats. The study investigated gene regulation in the mammalian eye and aimed to identify genetic variation relevant to human eye disease.

This dataset comprises expression values of 31042 probes on 120 twelve-week-old male offspring of rats. Following Huang, Ma, and Zhang (2008), we preprocess the data in two steps. First, we remove all probes for which the maximum expression among the 12 rats is less than the 25th percentile of the entire expression values. Second, we remove any probe for which the range of the expression among 120 rats is less than 2. After the preprocessing, 18,958 probes remain. We are interested in how the expression of gene TRIM32 (a gene identified to be associated with human hereditary diseases of the retina), corresponding to probe 1389163_at, depends on the expression values at other probes.

We rank all remaining 18,958 probes according to the absolute value of the correlations of their expression to the expression of probe 1389163_at and select the top 3000 probes. On this subset ($n = 120$, $p = 3000$), we apply the QICD algorithm to investigate the relationship between the expression of TRIM32 and those of the 3000 genes. First, we analyze the data on all 120 rats using SCAD or MCP penalized quantile regression ($\tau = 0.3$, 0.5, and 0.7). Again, we use the HBIC to select the tuning parameter λ for each case. The second column of Table 3 reports the number of nonzero coefficients (# nonzero) selected in each case. An interesting observation is that different sets of probes are selected at different quantiles. For example, for the SCAD penalty, the 18 probes selected at $\tau = 0.3$ and the 21 probe sets selected at $\tau = 0.7$ share 6 common ones ("1368887_at," "1382291_at," "1390048_at," "1380371_at," "1395973_at," "1374786_at"). However, none of six probes is selected at $\tau = 0.5$. A similar phenomenon was found for the MCP penalty. This suggests the presence of heterogeneity in the data.

We also randomly partition the 120 rats 50 times. In each partition, we randomly select 80 rats for the training set and have the rest for the test set. We fit penalized quantile regression model and select the tuning parameter on the training set. Then, we evaluate the predictive performance of the selected model using the test set. In the third column of Table 3, we report the average number of nonzero regression coefficients of the selected model and their associated robust standard deviations over the 50 repetitions. The last column of Table 3 contains the prediction error (and its standard deviation) on the test data, where

Table 3. Analysis of microarray dataset

	All data	Random partition	
Method	# nonzero	ave # nonzero	Prediction error
QICD-SCAD ($\tau = 0.5$)	18	16.53 (6.59)	1.72 (0.22)
QICD-SCAD ($\tau = 0.3$)	21	18.08 (7.09)	1.57 (0.28)
QICD-SCAD ($\tau = 0.7$)	13	12.06 (6.63)	1.53 (0.19)
QICD-MCP ($\tau = 0.5$)	19	15.61 (6.32)	1.73 (0.22)
QICD-MCP ($\tau = 0.3$)	23	16.86 (6.60)	1.56 (0.26)
QICD-MCP ($\tau = 0.7$)	12	11.12 (4.87)	1.52 (0.18)

the predication error is computed using the quantile loss function at the corresponding τ, that is, $\sum_{i=1}^{40} \rho_\tau(y_i - \hat{y}_i)$. We observe that the performance for the SCAD penalty and the MCP penalty is similar. At quantiles $\tau = 0.5$ and 0.7, we tend to select fewer probes than at $\tau = 0.3$. At $\tau = 0.3$ and 0.7, we observe a smaller prediction error than that at $\tau = 0.5$.

6. DISCUSSIONS

The article makes two timely contributions to nonconvex penalized quantile regression analysis of high-dimensional data. It proposes a fast iterative coordinate descent algorithm which significantly improves the computational speed for large p. Furthermore, we extend Tseng's (2001) theory to establish the convergence property of the proposed algorithm.

It is noted that although the majorization step is adopted, this step alone does not lead to the desirable computational efficiency gain for nonconvex penalized quantile regression. A stable and versatile class of MM algorithms applicable to a wide variety of penalization problems with nonconvex penalty was given in Schifano, Strawderman, and Wells (2010). They established a local convergence theory but required a strict convexity condition, which excludes the more interesting $p > n$ case.

APPENDIX A: SOME RELEVANT NOTATION AND RESULTS FROM TSENG (2001)

Let $\mathbf{R}^m$ denote the m-dimensional real space. For any $h : \mathbf{R}^m \mapsto \mathbf{R} \bigcup \infty$, we denote by dom h the effective domain of h, that is,

$$\text{dom } h = \{\mathbf{x} \in \mathbf{R}^m | h(\mathbf{x}) < \infty\}$$

For any $\mathbf{x} \in \text{dom} h$ and any $d \in \mathbf{R}^m$, we denote the *(lower) directional derivative* of h at $\mathbf{x}$ in the direction d by

$$h'(\mathbf{x}; d) = \liminf_{\lambda \downarrow 0} [h(\mathbf{x} + \lambda d) - h(\mathbf{x})]/\lambda.$$

We say that h is *quasiconvex* if

$$h(\mathbf{x} + \lambda d) \leq \max\{h(\mathbf{x}), h(\mathbf{x} + d)\},$$

and that h is *hemivariate* if h is not constant on any line segment belonging to dom h. A function f is said to be *lower semicontinuous (lsc)* on dom f if

$$\liminf_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0), \quad \text{for each } \mathbf{x}_0 \in \text{dom} f.$$

We define $\mathbf{z}$ as a *stationary point* of f if $\mathbf{z} \in \text{dom} f$ and

$$f'(\mathbf{z}; d) \geq 0, \quad \forall\ d.$$

We say that $\mathbf{z}$ is a *coordinate-wise minimum point* of f if $\mathbf{z} \in \text{dom} f$ and

$$f(\mathbf{z} + (0, \ldots, d_k, \ldots, 0)) \geq f(\mathbf{z}), \quad \forall\ d_k \in \mathbf{R}.$$

for all $k = 1, \ldots, N$. Here, we denote by $(0, \ldots, d_k, \ldots, 0)$ the vector in $\mathbf{R}^N$ whose kth coordinate is d_k and whose other coordinates are zero. We say that f is *regular* at $\mathbf{z} \in \text{dom } f$ if

$$f'(\mathbf{z}; d) \geq 0, \quad \forall d = (d_1, \ldots, d_N),$$
$$if f'(\mathbf{z}; (0, \ldots, d_k, \ldots, 0)) \geq 0, k = 1, \ldots, N. \tag{A.1}$$

We consider a generalized penalized loss function of the form

$$f(x_1, \ldots, x_N) = f_0(x_1, \ldots, x_N) + \sum_{k=1}^{N} f_k(x_k) \tag{A.2}$$

where $f_0 : \mathbf{R}^N \mapsto \mathbf{R} \bigcup \infty$ and $f_k : \mathbf{R} \mapsto \mathbf{R} \bigcup \infty$, $k = 1, 2, \ldots, N$, with We assume that f is proper, that is, $f \not\equiv \infty$. We adopt the following assumptions on $f, f_0, f_1, \ldots, f_N$.
(A1) f_0 is continuous on dom f_0.
(A2) For each $k \in \{1, \ldots, N\}$ and $(x_j)_{j \neq k}$, the function $x_k \mapsto f(x_1, \ldots, x_N)$ is *quasiconvex* and *hemivariate*.
(A3) $f_0, f_1, \ldots, f_N$ is *lower semicontinuous*.
(A4) dom f_0 is open and f_0 tends to ∞ at every boundary point of dom f_0.
(A5) dom $f_0 = Y_1 \times \ldots \times Y_N$, for some $Y_k \subseteq \mathbf{R}$, $k = 1, \ldots, N$.

In the following, we state a useful result of Tseng (2001).

Proposition A.1. (Tseng 2001) Consider an objective function of the form (A.2). Assume that $f, f_0, f_1, \ldots, f_N$ satisfy conditions (A1)–(A3) and that f_0 satisfies either condition (A4) or (A5). Let $\mathbf{x}^r = (x_1^r, \ldots, x_N^r)_{r=0,1,\ldots}$ be a sequence generated by the coordinate descent algorithm for minimizing (A.2) using the cyclic rule such as the one in (9). Then, either $\{f(\mathbf{x}^r)\} \downarrow -\infty$, or every cluster point $\mathbf{z}$ of $\{\mathbf{x}^r\}$ is a coordinatewise minimum point of f.

APPENDIX B: PROOF OF THE CONVERGENCE OF THE ALGORITHM

Proof of Lemma 1.

1. It is easy to show that $\phi_{\beta_j^{(k-1)}}\left(\beta_j^{(k-1)}\right) = p_\lambda\left(|\beta_j^{(k-1)}|\right)$. If $|\beta| > |\beta_j^{(k-1)}|$, by the mean value theorem, we have

$$p_\lambda(|\beta|) - p_\lambda\left(|\beta_j^{(k-1)}|\right) = p'_\lambda(\xi+)\left(|\beta| - |\beta_j^{(k-1)}|\right)$$

for some $\xi \in [|\beta_j^{(k-1)}|, |\beta|]$. Since $p_\lambda(\cdot)$ is concave, we have $p_\lambda{}'(|\beta_j^{(k-1)}|+) \geq p_\lambda{}'(\xi+)$. Hence,

$$\begin{aligned} p_\lambda(|\beta|) &= p_\lambda\left(|\beta_j^{(k-1)}|\right) + p'_\lambda(\xi+)\left(|\beta| - |\beta_j^{(k-1)}|\right) \\ &\leq p_\lambda\left(|\beta_j^{(k-1)}|\right) + p'_\lambda\left(|\beta_j^{(k-1)}|+\right)\left(|\beta| - |\beta_j^{(k-1)}|\right) \end{aligned}$$

$$= \phi_{\beta_j^{(k-1)}}(\beta).$$

Similarly, we can show that if $|\beta| < |\beta_j^{(k-1)}|$, then $p_\lambda(|\beta|) \le \phi_{\beta_j^{(k-1)}}(\beta)$. Therefore, $\phi_{\beta_j^{(k-1)}}(\beta)$ *majorizes* $p_\lambda(\beta)$ at the points $\pm|\beta_j^{(k-1)}|$.

2. It follows from 1 that

$$\begin{aligned} Q(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \\ &\le n^{-1}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) + \sum_{j=1}^{p} \phi_{\beta_j^{(k-1)}}(|\beta_j|) \\ &= Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}) \end{aligned}$$

Hence, $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ majorizes $Q(\boldsymbol{\beta})$ at the points $\pm|\boldsymbol{\beta}^{(k-1)}|$.

3. We have

$$Q(\boldsymbol{\beta}^{(k)}) \le Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k)}) \le Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k-1)}) = Q(\boldsymbol{\beta}^{(k-1)}),$$

where the first inequality and the last equality follow from the property of the majorization function in (4), while the second inequality follows from (8). This proves the descent property.

Proof of Lemma 2. The result follows directly from Proposition A.1. It is easy to check that $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ has the form (A.2) with components $f_0 = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})$ and $f_j = \phi_{\beta_j^{(k-1)}}(|\beta_j|)$ for $i \ge 1$ and $j \ge 1$, which satisfy conditions (A1)–(A5). Our algorithm implies that $\boldsymbol{\beta}^{(k)}$ is a cluster point of $\boldsymbol{\beta}_j^{(k)(r)}$. In addition, $Q(\boldsymbol{\beta}^{(k)}) \le Q(\boldsymbol{\beta}^{(0)}) < +\infty$, and $Q_{\boldsymbol{\beta}^{(k-1)}}(\beta) \ge 0$, $\{Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}_j^{(k)(r)})\} \not\to -\infty$ as $r \to \infty$. Hence, by Proposition A.1, $\boldsymbol{\beta}^{(k)}$ is a *coordinatewise minimum point* of $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$.

Proof of Lemma 3. We have

$$\begin{aligned} \lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k)}) &\ge \lim_{k\to+\infty} Q(\boldsymbol{\beta}^{(k)}) \qquad \text{(due to majorization)}. \\ &= \lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k)}) \\ &\ge \lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)}), \end{aligned}$$

where the last inequality follows because $\boldsymbol{\beta}^{(k)}$ is a coordinate minimum point. Hence $Q(\boldsymbol{\beta}^*) = A$.

To prove Theorem 1, we state below a useful result from Bazaraa, Sherali, and Shetty (2006, Theorem 3.3.10).

Lemma B.1. (Bazaraa, Sherali, and Shetty 2006) Given a function $f\colon \mathbf{R}^n \mapsto \mathbf{R}$, let $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda) = f(\bar{\mathbf{x}} + \lambda\mathbf{d})$, where $\bar{\mathbf{x}}$ is some point in $\mathbf{R}^n$ and $\mathbf{d} \in \mathbf{R}^n$ is a nonezero direction.

Then, f is (strictly) convex if and only if $F_{(\bar{\mathbf{x}};\mathbf{d})}(\cdot)$ is a (strictly) convex function of λ for all $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbf{R}^n$.

Proof. We include the proof here for completeness, Given any $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbf{R}^n$, we write $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda)$ as $F(\lambda)$ for notational simplicity. If f is convex, then for any λ_1 and λ_2 in $\mathbf{R}$ and for any $0 \leq \alpha \leq 1$, we have

$$\begin{aligned} F(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= f(\alpha[\bar{\mathbf{x}} + \lambda_1\mathbf{d}] + (1-\alpha)[\bar{\mathbf{x}} + \lambda_2\mathbf{d}]) \\ &\leq \alpha f(\bar{\mathbf{x}} + \lambda_1\mathbf{d}) + (1-\alpha) f(\bar{\mathbf{x}} + \lambda_2\mathbf{d}) \\ &= \alpha F(\lambda_1) + (1-\alpha) F(\lambda_2) \end{aligned}$$

Hence, F is convex. Conversely, suppose that $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda)$, $\lambda \in \mathbf{R}$, is convex for all $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbf{R}^n$. Then, for any $\mathbf{x}_1$ and $\mathbf{x}_2$ and $0 \leq \lambda \leq 1$, we have

$$\begin{aligned} \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) &= \lambda f[\mathbf{x}_1 + 0(\mathbf{x}_2 - \mathbf{x}_1)] + (1-\lambda) f[\mathbf{x}_1 + 1(\mathbf{x}_2 - \mathbf{x}_1)] \\ &= \lambda F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(0) + (1-\lambda) F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(1) \\ &\geq F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(1-\lambda) \\ &= f[\mathbf{x}_1 + (1-\lambda)(\mathbf{x}_2 - \mathbf{x}_1)] \\ &= f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2]. \end{aligned}$$

So f is convex. The argument for the strictly convex case is similar.

□

Proof of Theorem 1. Let $\{r_m\}$ be a subsequence of $\{k_m\}$ such that $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m)} = \boldsymbol{\beta}^*$ and $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m-1)} = \boldsymbol{\beta}^{**}$. Denote $\lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k)})$ by A. By Lemma 3, we have

$$\begin{aligned} Q(\boldsymbol{\beta}^*) = Q(\boldsymbol{\beta}^{**}) &= \lim_{m\to+\infty} Q_{\boldsymbol{\beta}^{(r_m-1)}}(\boldsymbol{\beta}^{(r_m)}) \\ &= Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*) = A. \end{aligned}$$

Note that $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$. By Lemma B.1, $R(\lambda) = Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta} + \lambda\mathbf{d})$ is convex in λ for all $\boldsymbol{\beta}$ and $\mathbf{d} \neq \mathbf{0}$. Moreover, by Lemma 2, $\boldsymbol{\beta}^{(r_m)}$ is the coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{(r_m-1)}}(\boldsymbol{\beta})$ for $m = 1, 2, \ldots$. Since $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m)} = \boldsymbol{\beta}^*$ and $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m-1)} = \boldsymbol{\beta}^{**}$, $\boldsymbol{\beta}^*$ is a coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ by the continuity of $Q_{\boldsymbol{\beta}^{(r_m-1)}}(\boldsymbol{\beta}^{(r_m)})$ and $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$. Since $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ is *regular* at $\boldsymbol{\beta}^*$, $\boldsymbol{\beta}^*$ is a stationary point as well by (A.1). Hence, we have

$$Q'_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*; \mathbf{d}) \geq 0, \quad \forall\, \mathbf{d},$$

which is equivalent to

$$R'(\lambda; \boldsymbol{\beta}^*, \mathbf{d})|_{\lambda=0} \geq 0, \quad \forall\, \mathbf{d}. \tag{B.1}$$

If $\boldsymbol{\beta}^{**}$ is not a coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$, then $\forall\ a > 0$, $\exists\ \mathbf{d}^{(1)} = (0, \ldots, d_i, \ldots, 0)$, with $|d_i| < a$, such that

$$Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**} + \mathbf{d}^{(1)}) < Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**}) = Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*) \tag{B.2}$$

Let $\mathbf{d}^{(2)} = \boldsymbol{\beta}^{**} + \mathbf{d}^{(1)} - \boldsymbol{\beta}^{*}$. Then, we have, $\forall\, \lambda \in (0, 1)$,

$$\begin{aligned} Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{*} + \lambda \mathbf{d}^{(2)}) &= Q_{\boldsymbol{\beta}^{**}}((1-\lambda)\boldsymbol{\beta}^{*} + \lambda(\boldsymbol{\beta}^{**} + \mathbf{d}^{(1)})) \\ &\leq (1-\lambda) Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{*}) + \lambda Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**} + \mathbf{d}^{(1)}) \\ &< Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{*}), \end{aligned}$$

where the last inequality follows from (B.2). Note that although $R(\lambda; \boldsymbol{\beta}^{*})$ is not differentiable everywhere, it is nondifferentiable at only a finite number of points; hence, there exits a constant λ, $R(\lambda; \boldsymbol{\beta}^{*})$ is differentiable in $(0, \lambda)$. Then, by the mean value theorem there exits $\lambda_1 \in (0, \lambda)$ such that

$$\begin{aligned} R'(\lambda_1; \boldsymbol{\beta}^{*}, \mathbf{d}^{(2)}) &= \frac{R(\lambda; \boldsymbol{\beta}^{*}, \mathbf{d}^{(2)}) - R(0; \boldsymbol{\beta}^{*}, \mathbf{d}^{(2)})}{\lambda} \\ &= \frac{Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{*} + \lambda \mathbf{d}^{(2)}) - Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{*})}{\lambda} \\ &< 0. \end{aligned}$$

However, $R(\cdot; \boldsymbol{\beta}^{*}, \mathbf{d}^{(2)})$ is convex. Hence, $R^{'}(0; \boldsymbol{\beta}^{*}, \mathbf{d}^{(2)}) \leq R'(\lambda_1; \boldsymbol{\beta}^{*}, \mathbf{d}^{(2)}) < 0$. This contradicts (B.1). Therefore, $\boldsymbol{\beta}^{**}$ is a coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$. Similarly, $\boldsymbol{\beta}^{**}$ is a stationary point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$. Furthermore, since $p'_{\lambda}(|\theta|+) = p'_{\lambda}(|\theta|-)$ on $(0, \infty)$, we have

$$Q'(\boldsymbol{\beta}^{**}; \mathbf{d}) = Q'_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**}; \mathbf{d}) \geq 0, \qquad \forall\, \mathbf{d}.$$

Hence, $\boldsymbol{\beta}^{**}$ is a *stationary point* of $Q(\boldsymbol{\beta})$. Since $\boldsymbol{\beta}^{**}$ is an arbitrary cluster point of $\{\boldsymbol{\beta}^{(k-1)}\}$, we conclude that every cluster point of the sequence generated by the QICD algorithm is a stationary point of $Q(\boldsymbol{\beta})$.

ACKNOWLEDGEMENT

This research is supported by NSF DMS-1308960. We thank the associate editor and two referees for their helpful comments. We also thank Yi Yang for sharing with us his codes on a faster approach for computing the weighted median, which we incorporated in our algorithm in the revision.

[Received April 2013. Revised March 2014.]

REFERENCES

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006), *Nonlinear Programming*, Hoboken, New Jersey: Wiley. [10,16]

Belloni, A., and Chernozhukov, V. (2011), "L1-Penalized Quantile Regression in High-Dimensional Sparse Models," *The Annals of Statistics*, 39, 82–130. [2]

Bickel, P., and Li, B. (2006), "Regularization in Statistics" (with discussion), *Sociedad de Estadística e Investigación Operativa Test*, 15, 271–344. [2]

Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection," *The Annals of Applied Statistics*, 5, 232–253. [3]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection With Large Model Space," *Biometrika*, 95, 759–771. [9]

Daubechies, I., Defrise, M., and De Mol, C. (2004), "An Iterative Thresholding Algorithm for Linear Inverse Problems With a Sparsity Constraint," *Communications on Pure and Applied Mathematics*, 57, 1413–1457. [3]

Fan, J., and Li, R. (2001), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642. [2,4]

Fan, J., and Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [2]

Friedman, J. H., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332. [3]

Fu, W. J. (1998), "Penalized Regressions: The Bridge Versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416. [3]

Huang, J., Ma, S., and Zhang, C. H. (2008), "Adaptive Lasso for Sparse High-Dimensional Regression Models," *Statistica Sinica*, 18, 1603–1618. [13]

Hunter, D. R., and Lange, K. (2000), "Quantile Regression via an MM Algorithm," *Journal of Computational and Graphical Statistics*, 9, 60–77. [2]

——— (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58, 30–37. [3]

Hunter, D. R., and Li, R. (2005a), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642. [3]

——— (2005b), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642. [xxxx]

Jiang, D., and Huang, J. (2012), "Majorization Minimization by Coordinate Descent for Concave Penalized Generalized Linear Models," Department of Biostatistics, University of Iowa, Report No. 412. 37p. [3,7]

Kai, B., Li, R., and Zou, H. (2011), "New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models," *The Annals of Statistics*, 39, 305–332. [2]

Kim, Y., Kwon, S., and Choi, H. (2012), "Consistent Model Selection Criteria on High Dimensions," *Journal of Machine Learning Research*, 13, 1037–1057. [9]

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press. [3]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [3]

Koenker, R., and Park, B. J. (1996), "An Interior Point Algorithm for Nonlinear Quantile Regression," *Journal of Econometrics*, 71, 265–283. [2]

Lange, K. (2004), *Optimization*, New York: Springer. [3]

Lee, E. R., Noh, H., and Park, B. U. (2014), "Model Selection via Bayesian Information Criterion for Quantile Regression Models," *Journal of the American Statistical Association*, 109, 216–229. [9]

Li, Y. J., and Zhu. J., (2008), "L_1-Norm Quantile Regression," *Journal of Computational and Graphical Statistics*, 17, 163–185. [2]

Mazumder, R., Friedman, J. H., and Hastie, T. (2011), "SparseNet: Coordinate Descent With NonConvex Penalties," *Journal of American Statistical Association*, 106, 1125–1138. [3]

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006), "Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease," *Proceedings of the National Academy of Sciences*, 103, 14429–14434. [13]

Schifano, E. D., Strawderman, R. L., and Wells, M. T. (2010), "Majorization-Minimization Algorithms for Nonsmoothly Penalized Objective Functions," *Electronic Journal of Statistics*, 4, 1258–1299. [14]

Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, 109, 475–494. [3,7,10,14,15]

Wang, H., Li, B., and Leng, C. (2009), "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters,," *Journal of Royal Statistical Society,* Series B, 71, 671–683. [9]

Wang, H., Li, R., and Tsai, C. L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [9]

Wang, L., Kim, Y. D., and Li, R. (2013), "Calibrating NonConvex Penalized Regression in Ultra-High Dimension," *The Annals of Statistics*, 41, 2505–2536. [9]

Wang, L., Wu, Y., and Li, R. (2012), "Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension," *Journal of American Statistical Association*, 107, 214–222. [2,5,11]

Wu, T. T., and Lange, K. (2008), "Coordinate Descent Algorithms for Lasso Penalized Regression," *The Annals of Applied Statistics*, 2, 224–244. [2,3,8]

Wu, Y., and Liu, Y. (2009), "Variable Selection in Quantile Regression," *Statistica Sinica*, 19, 801–817. [2]

Zhang, C. H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [2,4]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533. [2,5]

Zou, H., and Yuan, M. (2008), "Composite Quantile Regression and the Oracle Model Selection Theory," *The Annals of Statistics*, 36, 1108–1126. [2]